*The Annals of Statistics*
2020, Vol. 48, No. 2, 629–654
https://doi.org/10.1214/18-AOS1792


# ON ESTIMATION OF ISOTONIC PIECEWISE CONSTANT SIGNALS

BY CHAO GAO[1], FANG HAN[2] AND CUN-HUI ZHANG[3]

[1]*Department of Statistics, University of Chicago, chaogao@galton.uchicago.edu*
[2]*Department of Statistics, University of Washington, fanghan@uw.edu*
[3]*Department of Statistics and Biostatistics, Rutgers University, cunhui@stat.rutgers.edu*

Consider a sequence of real data points $X_1, \ldots, X_n$ with underlying means $\theta_1^*, \ldots, \theta_n^*$. This paper starts from studying the setting that $\theta_i^*$ is both piecewise constant and monotone as a function of the index $i$. For this, we establish the exact minimax rate of estimating such monotone functions, and thus give a nontrivial answer to an open problem in the shape-constrained analysis literature. The minimax rate under the loss of the sum of squared errors involves an interesting iterated logarithmic dependence on the dimension, a phenomenon that is revealed through characterizing the interplay between the isotonic shape constraint and model selection complexity. We then develop a penalized least-squares procedure for estimating the vector $\theta^* = (\theta_1^*, \ldots, \theta_n^*)^\mathsf{T}$. This estimator is shown to achieve the derived minimax rate adaptively. For the proposed estimator, we further allow the model to be misspecified and derive oracle inequalities with the optimal rates, and show there exists a computationally efficient algorithm to compute the exact solution.

**1. Introduction.** Consider an observed vector $X = (X_1, \ldots, X_n)^\mathsf{T}$ of independent entries and an unknown underlying mean $\theta^* = (\theta_1^*, \ldots, \theta_n^*)^\mathsf{T}$. This paper starts from the problem of estimating such $\theta^*$ that is isotonic piecewise constant. Specifically, for any $k \in (0 : n]$, we define the parameter space of interest as the set of all nondecreasing vectors with at most $k$ pieces,

$$\begin{aligned}\Theta_k^{\uparrow} = \big\{\theta \in \mathbb{R}^n : &\text{ there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that}\\ &0 = a_0 \le a_1 \le \cdots \le a_k = n,\\ &\mu_1 \le \mu_2 \le \cdots \le \mu_k, \text{ and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j]\big\}.\end{aligned}$$

The notation $(a : b]$ stands for the set of all integers $i$ that satisfy $a < i \le b$. For any vector $\theta^* \in \Theta_k^{\uparrow}$, it is a piecewise constant signal with at most $k$ steps that take different values. When $k = n$, the space $\Theta_k^{\uparrow}$ contains all vectors $\theta^*$ that satisfy $\theta_1^* \le \theta_2^* \le \cdots \le \theta_n^*$. Estimation of $\theta^*$ under this condition is recognized as isotonic regression. It has been one of the most popular and successful directions in the shape-constrained analysis literature. General discussions on relevant methods and theory can be found in Robertson, Wright and Dykstra (1988), Groeneboom and Wellner (1992), Silvapulle and Sen (2011), and Groeneboom and Jongbloed (2014), to name just a few. However, in certain cases, isotonic regression may overfit the data by producing a result with too many steps. This inspires research on fitting isotonic regression with the restriction of the number of steps. According to Schell and Singh (1997), the problem is termed as reduced isotonic regression. The parameter space $\Theta_k^{\uparrow}$ precisely describes such regression functions.

Received July 2018; revised November 2018.
*MSC2010 subject classifications.* 62G08, 62C20.
*Key words and phrases.* Isotonic piecewise constant function, reduced isotonic regression, iterated logarithmic dependence, adaptive estimation, oracle inequalities.

Despite its practical importance in changepoint and shape-constrained analyses, the fundamental limit of estimating $\theta^*$ in the class $\Theta_k^{\uparrow}$ is still unknown. We summarize the results in the literature by assuming that $X \sim N(\theta^*, \sigma^2 I_n)$. In terms of upper bound, Chatterjee, Guntuboyina and Sen (2015) show explicitly that

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \leq C\sigma^2 k \log(en/k),$$

and the rate $\sigma^2 k \log(en/k)$ can be adaptively achieved by isotonic regression. See Bellec (2018) and Bellec and Tsybakov (2015) for results with the same rate. In terms of lower bound, Bellec and Tsybakov (2015) show

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \geq c\sigma^2 k.$$

We can see the above upper and lower bounds do not match, and it is unclear if either bound is sharp.

In this paper, we settle a solution to this open problem by deriving the precise minimax rate of the space $\Theta_k^{\uparrow}$. Thus, the gap between the upper and lower bounds in the literature is closed. Surprisingly, neither the upper nor the lower bound in the literature is sharp. We prove that for $k \geq 2$, the minimax rate takes the form

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \asymp \sigma^2 k \log\log(16n/k).$$

It is interesting that the minimax rate of the problem has an iterated logarithmic dependence on $n/k$, an engaging feature of the space $\Theta_k^{\uparrow}$.

We show that the minimax rate can be achieved by solving a least-squares problem in the space $\Theta_k^{\uparrow}$. This is exactly the procedure of reduced isotonic regression. In comparison, the ordinary isotonic regression proves to achieve only a suboptimal rate $\sigma^2 k \log(en/k)$. Therefore, our results provide a theoretical justification that the reduced isotonic regression can avoid overfitting the data and practically attain better performances over the ordinary isotonic regression (cf. Schell and Singh (1997), Salanti and Ulm (2003), Haiminen, Gionis and Laasonen (2008)).

The proof of the result is nontrivial. Our analysis involves repeatedly partitioning the studied sequence according to the nature of the reduced isotonic regression estimator. This allows us to use martingale maximal inequalities by Levy and Doob, and gives us the sharp minimax rate.

Besides understanding the fundamental challenge in estimating the piecewise monotone functions, in practice, it is always the case that: (i) the number of steps or pieces $k$ is unknown; (ii) the model could be misspecified. In addition, practically we would love to have a computationally feasible algorithm to compute the exact solution. Indeed, in this manuscript we propose a penalized least-squares (reduced isotonic regression) estimator that achieves the minimax rate without knowing $k$. We further allow the model to be misspecified and prove oracle inequalities with the optimal rates. Moreover, by exploring a key property of reduced isotonic regression and by leveraging the pool-adjacent-violators algorithm (PAVA) (Mair, Hornik and de Leeuw (2009)), we develop a computationally efficient algorithm to compute the $k$-piece least-squares estimator for all $k$ and thus the penalized least-squares estimator.

This paper also obtains exact minimax rates under the $\ell_p$ loss with $1 \leq p < 2$. In contrast to the case $p = 2$, the minimax rates are now parametric. Furthermore, we show that this rate can be adaptively achieved by isotonic regression, but not by the reduced isotonic regression procedure. In other words, the nature of the problem can be dramatically changed by using a different loss function.

The rest of the paper is organized as follows. In Section 2, we introduce the problem setting and present the minimax rate. We then introduce an adaptive estimation procedure in Section 3. The computational issues of the estimators are discussed in Section 4. We will also put our results in a larger picture and discuss a few other related problems in Section 4. All the proofs are relegated to Section 5 and the Supplementary Material (Gao, Han and Zhang (2020)).

*Notation.* Let $\mathbb{Z}$ and $\mathbb{R}$ be the sets of integers and real numbers. For any positive integer $d$, we use $[d]$ to denote the set $\{1, 2, \ldots, d\}$. Let $\mathbb{1}(\cdot)$ denote the indicator function. For a real number $x$, $\lceil x \rceil$ is the smallest integer no smaller than $x$, $\lfloor x \rfloor$ is the largest integer no larger than $x$, $x_+ = x\mathbb{1}(x \geq 0)$ and $x_- = -x\mathbb{1}(x < 0)$ are the positive and negative components of $x$. For any $a, b \in \mathbb{R}$, write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For an arbitrary vector $\theta = (\theta_1, \ldots, \theta_n)^\mathsf{T} \in \mathbb{R}^n$ and an index set $J \subset [n]$, we denote $\theta_J$ to be the subvector of $\theta$ with entries indexed by $J$, and for any $p \geq 1$,

$$\|\theta\|_p = \left(\sum_{i=1}^n |\theta_i|^p\right)^{1/p}, \quad \text{and} \quad \|\theta\|_{J,p} = \left(\sum_{i \in J} \theta_i^p\right)^{1/p}.$$

In particular, we denote $\|\theta\| = \|\theta\|_2$ and $\|\theta\|_J = \|\theta\|_{J,2}$. Let $\overline{\theta}_J = \frac{1}{|J|}\sum_{i \in J} \theta_i$ represent the sample mean across the sequence $\theta_J$. For any real value $a$ and positive integer $n$, define

$$\{a\}^n = (\underbrace{a, a, \ldots, a}_{n})^\mathsf{T}.$$

For any sets of vectors $\Theta_1 \subset \mathbb{R}^{n_1}, \ldots, \Theta_m \subset \mathbb{R}^{n_m}$, denote

$$\bigtimes_{\ell=1}^m \Theta_\ell = \{\theta = (\theta_{(1)}^\mathsf{T}, \ldots, \theta_{(m)}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{\sum_{i=1}^m n_i} : \theta_{(\ell)} \in \Theta_\ell\}.$$

Throughout the paper, let $c, C, c_1, C_1, c_2, C_2, \ldots$ be generic universal positive constants whose actual values may vary at different places. For any two positive data sequences $\{a_n, n = 1, 2, \ldots\}$ and $\{b_n, n = 1, 2, \ldots\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all $n$ from natural numbers. The notation $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation operations whenever the distributions can be determined from the context.

**2. Minimax rates.** In this section, we present the minimax rate of the space $\Theta_k^\uparrow$ with respect to the squared $\ell_2$ loss. We first consider the upper bound. Given the observation $X \in \mathbb{R}^n$, we define the constrained least-squares estimator as

$$\widehat{\theta}(\Theta_k^\uparrow) = \operatorname*{argmin}_{\theta \in \Theta_k^\uparrow} \|X - \theta\|^2. \tag{1}$$

Computational issues related to this estimator will be discussed in Section 4.1. Note that if $X \sim N(\theta^*, \sigma^2 I_n)$, $\widehat{\theta}(\Theta_k^\uparrow)$ is simply the maximum likelihood estimator (MLE) restricted onto the parameter space $\Theta_k^\uparrow$. However, we do not need to assume a Gaussian error for the risk bound presented below. In detail, consider the observation

$$X = \theta^* + Z,$$

where we assume the error variables $\{Z_i\}_{i=1}^n$ are independent with zero mean and satisfy one of the following conditions:

$$\begin{cases} \max\limits_{1 \leq i \leq n} \mathbb{E}|Z_i/\sigma|^{2+\epsilon} \leq C_1 & \text{not identically distributed } Z_i\text{'s}, \\ \mathbb{E}(Z_1^2/\sigma^2)\log(e + Z_1^2/\sigma^2) \leq C_1 & \text{identically distributed } Z_i\text{'s}, \end{cases} \tag{2}$$

for some number $\sigma > 0$, an arbitrarily small universal constant $\epsilon \in (0, 1)$, and some universal constant $C_1 > 0$. It is easy to see that the Gaussian error $Z \sim N(0, \sigma^2 I_n)$ is a special case.

THEOREM 2.1. *Consider $X = \theta^* + Z$ with any $\theta^* \in \mathbb{R}^n$ and $Z$ satisfying (2). Then, we have*

$$\mathbb{E}\big\|\widehat{\theta}\big(\Theta_k^{\uparrow}\big) - \theta^*\big\|^2 \le C\Big[\inf_{\theta \in \Theta_k^{\uparrow}} \big\|\theta - \theta^*\big\|^2 + \sigma^2 + \sigma^2 k \log\log(16n/k)\mathbb{1}\{k \ge 2\}\Big]$$

*for all $k \in [n]$ with some universal constant $C > 0$.*

Note that Theorem 2.1 is an oracle inequality without any assumption on the true mean vector $\theta^*$. Besides the trivial bound $C(\inf_{\theta \in \Theta_1^{\uparrow}} \|\theta - \theta^*\|^2 + \sigma^2)$ for $k = 1$, it is interesting that the stochastic error scales as $\sigma^2 k \log\log(16n/k)$ for $k \ge 2$. This iterated logarithmic term appears due to the isotonic constraint of the solution $\widehat{\theta}(\Theta_k^{\uparrow})$ as well as the properties of partial sum processes. More technical discussions on this point will be given in Section 4.2, which discusses the importance of the isotonic constraint in more details.

If the condition $\theta^* \in \Theta_k^{\uparrow}$ holds, then we immediately obtain the following corollary:

$$\sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\widehat{\theta}\big(\Theta_k^{\uparrow}\big) - \theta^*\big\|^2 \le C\sigma^2 k \log\log(16n/k),$$

when $k \ge 2$. This improves previous risk bounds for the space $\theta^* \in \Theta_k^{\uparrow}$ in the literature. For example, for the ordinary isotonic regression estimator

$$\widehat{\theta}^{(\mathrm{iso})} = \widehat{\theta}\big(\Theta_n^{\uparrow}\big) = \operatorname*{argmin}_{\theta:\theta_1 \le \theta_2 \le \cdots \le \theta_n} \|X - \theta\|^2, \tag{3}$$

Theorem 2.1 of Zhang (2002) gives

$$\sum_{i=n_1+1}^{n_2} |\widehat{\theta}_i^{(\mathrm{iso})} - \theta_i^*|^2 \le \int_0^{n_2-n_1} \frac{C\sigma^2}{1 \vee x}\, dx,$$

whenever $0 \le n_1 < n_2 \le n$ and $\theta_{n_2}^* = \theta_{n_1+1}^*$ for a nondecreasing $\theta^*$. Thus, as explicitly derived in Chatterjee, Guntuboyina and Sen (2015),

$$\sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\widehat{\theta}^{(\mathrm{iso})} - \theta^*\big\|^2 \le C\sigma^2 k \log(en/k).$$

Our result shows that the logarithmic error term in the above bound can be improved by restricting the least-squares optimization to the space $\theta^* \in \Theta_k^{\uparrow}$. This modification of the estimator is necessary, as shown below.

PROPOSITION 2.1. *There exists a universal constant $c > 0$, such that*

$$\sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\widehat{\theta}^{(\mathrm{iso})} - \theta^*\big\|^2 \ge c\sigma^2 k \log(en/k).$$

Next, we show that the rate obtained by Theorem 2.1 is optimal by giving a matching minimax lower bound. To this end, we consider the Gaussian distribution $X \sim N(\theta^*, \sigma^2 I_n)$. In the following a lower bound construction for $k = 2$ is provided, with the generalization to $k \ge 2$ briefly sketched.

By Fano's inequality (Proposition 5.1), we need to find some subset $T \subset \Theta_2^{\uparrow}$ such that the ratio

$$\frac{\max_{\theta,\theta' \in T} \|\theta - \theta'\|^2/(2\sigma^2)}{\log \mathcal{M}(\epsilon, T, \|\cdot\|)}$$

is bounded by a sufficiently small constant. Here, $\mathcal{M}(\epsilon, T, \|\cdot\|)$ stands for the packing number of $T$ with radius $\epsilon$ and distance $\|\cdot\|$. We will take $\epsilon^2 \asymp \log\log(16n)$. Since the minimax rate is simply $\sigma^2$ if $n$ is bounded by a constant, we only need to construct $T$ with a sufficiently large $n$. For each $\ell \in \{1, 2, \ldots, \lceil \log_2 n \rceil\}$, construct the vector $\theta_\ell \in \mathbb{R}^n$ by filling the last $\lceil n2^{-\ell} \rceil$ entires with $\sqrt{\alpha\sigma^2 2^\ell \log\log_2 n/n}$ and the remaining entries 0. It is easy to see that $\theta_\ell \in \Theta_2^{\uparrow}$ for all $\ell \in \{1, 2, \ldots, \lceil \log_2 n \rceil\}$. For any $j < \ell$, we have

$$\begin{aligned}
\|\theta_\ell - \theta_j\|^2 &\geq \lceil n2^{-\ell} \rceil \left( \sqrt{\frac{\alpha\sigma^2 2^\ell \log\log_2 n}{n}} - \sqrt{\frac{\alpha\sigma^2 2^j \log\log_2 n}{n}} \right)^2 \\
&\geq \alpha\sigma^2 \log\log_2 n \left(1 - 2^{\frac{j-\ell}{2}}\right)^2 \\
&\geq \frac{\alpha\sigma^2}{20} \log\log_2 n.
\end{aligned}$$

Therefore,

$$\log \mathcal{M}\left( \sqrt{\frac{\alpha\sigma^2}{20} \log\log_2 n}, T, \|\cdot\| \right) \geq \log\log_2 n, \tag{4}$$

where $T = \{\theta_\ell : \ell = 1, 2, \ldots, \lceil \log_2 n \rceil\}$. Moreover, since $\|\theta_\ell\|^2 \leq 3\alpha\sigma^2 \log\log_2 n$ for all $\ell$, we have

$$\max_{\theta,\theta' \in T} \frac{1}{2\sigma^2} \|\theta - \theta'\|^2 \leq 6\alpha \log\log_2 n. \tag{5}$$

Hence, by (4) and (5), we can choose a very small $\alpha > 0$ to ensure the ratio $\frac{\max_{\theta,\theta' \in T} \|\theta-\theta'\|^2/(2\sigma^2)}{\log \mathcal{M}(\epsilon, T, \|\cdot\|)}$ to be small. This leads to the minimax lower bound

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_2^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta\|^2 \geq c\sigma^2 \log\log(16n),$$

for $k = 2$.

For a general $k > 2$, the idea is to divide the integer set $[n]$ into $\lceil k/2 \rceil - 1$ consecutive intervals with length approximately $\lfloor 2n/k \rfloor$. Then, we can apply the above construction to each of the $\lceil k/2 \rceil - 1$ interval. For each interval, a lower bound $c\sigma^2 \log\log(2n/k)$ is obtained. Summing up these lower bounds over all the $k/2$ intervals, we get the desired rate. Details of this argument will be given in Section 5.3, and the according minimax lower bound is presented as follows.

THEOREM 2.2. *There exists some universal constant $c > 0$, such that*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \geq \begin{cases} c\sigma^2, & k = 1, \\ c\sigma^2 k \log\log(16n/k), & k \geq 2, \end{cases}$$

*where the infimum is taken over all measurable functions of $X$ and the expectation is taken under which $X \sim N(\theta^*, \sigma^2 I_n)$.*

Combining the results of Theorem 2.1 and Theorem 2.2, we obtain the minimax rate of the problem

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \asymp \begin{cases} \sigma^2, & k = 1, \\ \sigma^2 k \log\log(16n/k), & 2 \le k \le n. \end{cases}$$

The minimax rate implies that the iterated logarithmic dependence on $n$ is an essential feature of the space $\Theta_k^{\uparrow}$.

**3. Adaptive estimation.** The estimator (1) that achieves the minimax rate requires the knowledge of $k$. This section proposes an adaptive estimator that can also achieve the minimax rate without knowing the value of $k$. Recalling the notation $\widehat{\theta}(\Theta_k^{\uparrow}) = \operatorname{argmin}_{\theta \in \Theta_k^{\uparrow}} \|X - \theta\|^2$, we propose an adaptive estimator $\widehat{\theta} = \widehat{\theta}(\Theta_{\widehat{k}}^{\uparrow})$ with a data-driven $\widehat{k}$. The data-driven $\widehat{k}$ is defined through the following penalized least-squares optimization. That is,

$$\widehat{k} = \operatorname*{argmin}_{k \in [n]} \big\{ \|X - \widehat{\theta}(\Theta_k^{\uparrow})\|^2 + \operatorname{pen}_{\tau}(k) \big\}. \tag{6}$$

Inspired by the minimax rate, the penalty function is defined by

$$\operatorname{pen}_{\tau}(k) = \begin{cases} \tau, & k = 1, \\ \tau k \log\log(16n/k), & 2 \le k \le n. \end{cases} \tag{7}$$

The estimator $\widehat{\theta}$ enjoys the following adaptive oracle inequality.

THEOREM 3.1. *Consider $X = \theta^* + Z$ with any $\theta^* \in \mathbb{R}^n$ and $Z$ satisfying (2). We use the estimator $\widehat{\theta} = \widehat{\theta}(\Theta_{\widehat{k}}^{\uparrow})$ with $\widehat{k}$ defined in (6). The tuning parameter is chosen as $\tau = C'\sigma^2$ for some sufficiently large universal constant $C' > 0$. Then, we have*

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \le C \min_{1 \le k \le n} \Big\{ \inf_{\theta \in \Theta_k^{\uparrow}} \|\theta - \theta^*\|^2 + \operatorname{pen}_{\tau}(k) \Big\}$$

*with some universal constant $C > 0$.*

REMARK 3.1. Unlike in isotonic regression, an implicit assumption of Theorem 3.1 is that we need to know the order of the variance $\sigma^2$. When $Z_i \sim N(0, \sigma^2)$, the unknown $\sigma$ can be estimated by the following robust procedure:

$$\widehat{\sigma} = \frac{\operatorname{Median}(|X_{i+1} - X_i|, 1 \le i < n)}{\sqrt{2}\operatorname{Median}(|N(0,1)|)}.$$

As $|\{i : |\mathbb{E}[X_{i+1} - X_i]| > \epsilon_0 \sigma\}|$ is bounded by $k-1$ when $\theta^*$ has $k$ pieces and by $\|\theta^*\|_1/(\epsilon_0\sigma)$ in general, the above $\widehat{\sigma}$ is consistent when $\min(k, \|\theta^*\|_1/\sigma) = o(n)$ and is of the order $\sigma$ when $\min(k, \|\theta^*\|_1/\sigma) \le c_0 n$ for some fixed small enough constant $c_0 > 0$. On the other hand, estimation of $\sigma^2$, or even just its order, is impossible when $\theta^*$ is arbitrary. In this case, whether it is still possible to achieve the oracle inequality in Theorem 3.1 is an interesting open problem.

Theorem 3.1 can be viewed as an adaptive version of Theorem 2.1. The oracle inequality automatically selects the best $k$ that achieves the optimal bias-variance tradeoff. When the true mean vector $\theta^*$ does belong to the space $\Theta_k^{\uparrow}$, we have $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \lesssim \operatorname{pen}_{\tau}(k)$, and thus the minimax rate is achieved without the knowledge of $k$.

When $\theta^* \in \Theta_n^{\uparrow}$ so that it is isotonic, the above oracle inequality can be further improved. By Meyer and Woodroofe (2000) and Zhang (2002), as $\theta^*$ is isotonic, the estimator $\widehat{\theta}^{(\mathrm{iso})} = \widehat{\theta}(\Theta_n^{\uparrow})$ satisfies the risk bound

$$\mathbb{E}\|\widehat{\theta}^{(\mathrm{iso})} - \theta^*\|^2 \lesssim \sigma^2\{\log(en) + n^{1/3}(V(\theta^*)/\sigma)^{2/3}\}, \tag{8}$$

where $V(\theta^*) = \theta_n^* - \theta_1^*$ is the total variation of the vector $\theta^*$. This risk bound can be significantly smaller than $\mathrm{pen}_\tau(k)$ when $V(\theta^*)/\sigma$ is small and $k$ is large. This motivates us to modify the value of $\mathrm{pen}_\tau(n)$ to achieve the better rate between (8) and (7). A direct choice of the modified penalty is just the bound on the right-hand side of (8). However, this option depends on the value of $V(\theta^*)$, which may not be available in practice. Inspired by the risk analysis in Zhang (2002), we consider

$$\tau\biggl\{\log(en) + \sum_{\{\ell \geq 0 : 2^\ell \leq n/3\}} \frac{\widehat{l}_\tau(2^{\ell+1}) - \widehat{l}_\tau(2^\ell)}{2^{\ell+1}}\biggr\}, \tag{9}$$

where

$$\hat{l}_\tau(m) := \min\{n, 3m + m\sqrt{m+1}(\overline{X}_{[n-m:n-m/2)} - \overline{X}_{(1+m/2:1+m]})/\sqrt{\tau}\}.$$

Note that (9) is a data-driven estimate of the risk of $\widehat{\theta}^{(\mathrm{iso})}$. Then, we have a well-defined penalty function on $[n]$ by combining (7) and (9). The modified penalty function in summary is

$$\widetilde{\mathrm{pen}}_\tau(k) = \begin{cases} \tau, & k = 1, \\ \tau\,\mathrm{pen}_\tau(k), & 2 \leq k \leq n-1, \\ \tau\biggl\{\log(en) + \displaystyle\sum_{\{\ell \geq 0 : 2^\ell \leq n/3\}} \frac{\widehat{l}_\tau(2^{\ell+1}) - \widehat{l}_\tau(2^\ell)}{2^{\ell+1}}\biggr\}, & k = n. \end{cases}$$

With some appropriate choice of $\tau$, the performance of $\widehat{\theta} = \widehat{\theta}(\Theta_{\widehat{k}}^{\uparrow})$ is given by the following theorem.

THEOREM 3.2. *Consider $X = \theta^* + Z$ with any $\theta^* \in \Theta_n^{\uparrow}$ and $Z$ satisfying $\max_{1 \leq i \leq n} \mathbb{E}|Z_i/\sigma|^{2+\epsilon} \leq C_1$. We use the estimator $\widehat{\theta} = \widehat{\theta}(\Theta_{\widehat{k}}^{\uparrow})$ with $\widehat{k}$ selected by the modified penalty function $\widetilde{\mathrm{pen}}_\tau(k)$. The tuning parameter is chosen as $\tau = C'\sigma^2$ for some sufficiently large universal constant $C' > 0$. Then, we have*

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \leq C \min_{1 \leq k \leq n}\Bigl\{\inf_{\theta \in \Theta_k^{\uparrow}} \|\theta - \theta^*\|^2 + \mathrm{isoerr}_k(\theta^*)\Bigr\},$$

*for some universal constant $C > 0$. The stochastic error term $\mathrm{isoerr}_k(\theta^*)$ is defined by*

$$\mathrm{isoerr}_k(\theta^*) = \begin{cases} \sigma^2, & k = 1, \\ \sigma^2 \min\biggl\{k \log\log\Bigl(\dfrac{16n}{k}\Bigr), \log(en) + n^{1/3}\Bigl(\dfrac{V(\theta^*)}{\sigma}\Bigr)^{2/3}\biggr\}, & k \geq 2. \end{cases}$$

We remark that the rate in the above theorem is always no greater than that of Theorem 3.1. If we further impose the condition that $V(\theta^*)/\sigma \leq n^{1-\delta}$ for some universal constant $\delta \in$

(0, 1), the rate given by Theorem 3.2 can be summarized into three phases:

$$\mathrm{isoerr}_k(\theta^*) \asymp \begin{cases} \sigma^2, \quad k = 1, \\ \sigma^2 k \log\log(16n), \\ \qquad 2 \le k \le \dfrac{\log(en) + n^{1/3}(V(\theta^*)/\sigma)^{2/3}}{\log\log(16n)}, \\ \sigma^2\{\log(en) + n^{1/3}(V(\theta^*)/\sigma)^{2/3}\}, \\ \qquad k > \dfrac{\log(en) + n^{1/3}(V(\theta^*)/\sigma)^{2/3}}{\log\log(16n)}. \end{cases}$$

In other words, the adaptive estimator with the modified penalty can achieve both the minimax rates of the class $\Theta_k^{\uparrow}$ derived in this paper and the rate of isotonic regression in Meyer and Woodroofe (2000) and Zhang (2002).

An interesting open problem is whether it is possible to obtain sharp oracle inequalities with the constant before the approximation error to be exactly one. The counter example constructed by Rigollet and Tsybakov (2012) in a sparse linear regression setting seems to suggest that this task may be impossible for the penalized least-squares procedure considered in this paper.

## 4. Discussion.

4.1. *Computational issues*. The optimization problem (1) is recognized as reduced isotonic regression in the literature (Schell and Singh (1997)), and related $\ell_0$ optimization problems have been studied in literature (see, e.g., Friedrich et al. (2008) and Jewell and Witten (2018) among many others). As $k = n$, the solution to the isotonic regression problem, $\widehat{\theta}(\Theta_n^{\uparrow})$, can be computed efficiently in $O(n)$ time using the pool-adjacent-violators algorithm (PAVA) (Mair, Hornik and de Leeuw (2009)). Computation of $\widehat{\theta}(\Theta_k^{\uparrow})$ for $k = 1, 2, \ldots, n-1$ may seem to be combinatorial, but by taking advantage of the PAVA solution, it can be reduced to a simple dynamic programming.

In detail, denote the set of knots (change points) of $\widehat{\theta}(\Theta_k^{\uparrow})$ by $\widehat{A}_k$. The following two properties are immediate from Lemma 5.1 (that will be stated in Section 5.1):

1. For any $k \in [n]$, we have $\widehat{A}_k \subset \widehat{A}_n$;
2. For any $k \in [n]$, $\widehat{\theta}(\Theta_k^{\uparrow})$ is a piecewise constant function with knots in $\widehat{A}_k$. Moreover, each piece is a sample average of the $X_i$'s in that block.

The first property asserts that the knots of $\widehat{\theta}(\Theta_k^{\uparrow})$ are always contained in the solution of PAVA. The second property implies that $\widehat{\theta}(\Theta_k^{\uparrow})$ can be obtained by averaging consecutive entries of $\widehat{\theta}(\Theta_n^{\uparrow})$. Since $\widehat{\theta}(\Theta_n^{\uparrow})$ is already isotonic, one does not need to worry about the isotonic constraint anymore, and the only task is to find the best change points among $\widehat{A}_n$ that minimize the squared error loss. Therefore, one can first run PAVA and obtain a set of potential knots $\widehat{A}_n = \{t_j\}_{j=1}^{\widehat{n}}$. Then, the search for the knots of $\widehat{\theta}(\Theta_k^{\uparrow})$ in $\{t_j\}_{j=1}^{\widehat{n}}$ can be implemented efficiently through dynamic programming. Note that $\widehat{A}_k = \widehat{A}_n$ for all $k \ge \widehat{n}$, and we only need to find $\widehat{A}_k$ for $k < \widehat{n}$. Details of implementation are given in Algorithm 1 for completedness.

Since Algorithm 1 computes $\widehat{\theta}(\Theta_k^{\uparrow})$ for all $k$, one can directly use the results to obtain the adaptive estimator $\widehat{\theta} = \widehat{\theta}(\Theta_{\widehat{k}}^{\uparrow})$ via (6). By Friedrich et al. (2008), the complexity of Algorithm 1 is $O(\widehat{n}^3)$ after PAVA. Therefore, the overall complexity of (6) is $O(n + \widehat{n}^3)$. This leads to a worst-case complexity bound $O(n^3)$. However, since $\widehat{n}$ enjoys the rate $\sigma^2\{V/\sigma + \log(en) +$

**Algorithm 1:** Computation of $\widehat{A}_k$ for all $k < \widehat{n}$

**Input** : $\{X_i\}_{i=1}^n$, $t_0 = 0$, knots $t_1 < \cdots < t_{\widehat{n}} = n$ from PAVA
**Output**: $\widehat{A}_k$ and the corresponding piecewise average for all $k < \widehat{n}$

**1** For $j$ in $1 : \widehat{n}$, compute the partial sums of $X_i$ and $X_i^2$,
$S(j) \leftarrow \sum_{0<i\le t_j} X_i$, $\quad \mathrm{SS}(j) \leftarrow \sum_{0<i\le t_j} X_i^2$.

**2** For all $(\ell, j)$ such that $0 \le \ell < j \le \widehat{n}$, compute the loss for fitting by mean in $(t_\ell : t_j]$,
$\mathrm{Loss}(\ell, j) \leftarrow \mathrm{SS}(j) - \mathrm{SS}(\ell) - (S(j) - S(\ell))^2/(a_j - a_\ell)$.

**3** For $j$ in $1 : \widehat{n}$, copy the loss for fitting by mean in $(0 : t_j]$,
$\mathrm{T.Loss}(1, j) \leftarrow \mathrm{Loss}(0, j)$.

**4** For $k$ in $2 : \widehat{n} - 1$
For $j$ in $k : \widehat{n}$, compute the minimal loss for $k$-piece monotone fit in $(0 : t_j]$,
$\mathrm{left.knot}(k, j) \leftarrow \mathrm{argmin}_{1\le\ell<j}\{\mathrm{T.Loss}(k-1, \ell) + \mathrm{Loss}(\ell, j)\}$,
$\mathrm{T.Loss}(k, j) \leftarrow \mathrm{T.Loss}(k-1, \mathrm{left.knot}(k, j)) + \mathrm{Loss}(\mathrm{left.knot}(k, j), j)$.
$\mathrm{knots}(k, k) = \widehat{n}$.
For $j$ in $(k-1) : 1$, compute $\widehat{A}_k$,
$\mathrm{knots}(k, j) \leftarrow \mathrm{left.knot}(j+1, \mathrm{knots}(k, j+1))$.

$n^{1/3}(V/\sigma)^{2/3}\}$ by Theorem 1 of Meyer and Woodroofe (2000), with high probability the isotonic regression (or PAVA) yields an $\widehat{n}$ of order $O(n^{1/3})$ when $V/\sigma = O(1)$. This leads to a linear complexity $O(n)$.

4.2. *Comparison with piecewise constant models.* A closely related problem to estimating isotonic piecewise constant functions is the estimation of piecewise constant signals without the monotone condition. We define the space of piecewise constant models as

$$\Theta_k = \big\{\theta \in \mathbb{R}^n : \text{there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that } 0 = a_0 \le a_1 \le \cdots \le a_k = n, \text{ and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j]\big\}. \tag{10}$$

This section shows that $\Theta_k^{\uparrow}$ and $\Theta_k$ have different error behaviors.

THEOREM 4.1. *For any $k \in [n]$, the minimax rate for the space $\Theta_k$ is given by*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \asymp \begin{cases} \sigma^2, & k = 1, \\ \sigma^2 \log\log(16n), & k = 2, \\ \sigma^2 k \log(en/k), & k \ge 3, \end{cases}$$

*where the expectation is taken over the distribution $X \sim N(\theta^*, \sigma^2 I_n)$.*

The upper bound in Theorem 4.1 can be achieved by the least-squares estimator $\widehat{\theta}(\Theta_k) = \mathrm{argmin}_{\theta\in\Theta_k} \|X - \theta^*\|^2$ when $k$ is known, or achieved by its penalized version when $k$ is unknown. The penalty can be chosen proportional to the minimax rate, following the classic approach in, for example, Birgé and Massart (1993) and Birgé and Massart (2001). These estimators can be computed efficiently via dynamic programming (Friedrich et al. (2008)).

We emphasize that the results for $k \ge 3$ are well known in the literature (Birgé and Massart (2001), Boysen et al. (2009), Donoho and Johnstone (1994), Li, Munk and Sieling (2016), Raskutti, Wainwright and Yu (2011)) and we claim no originality there. Instead, our stress is on comparing $\Theta_k^{\uparrow}$ and $\Theta_k$. First, it can be seen that the main difference between these two spaces is that the minimax rate of the former scales as $\sigma^2 k \log\log(16n/k)$, while that of the latter scales as $\sigma^2 k \log(en/k)$, for $k \ge 3$. The case $k = 2$ is special, and both spaces have

minimax rates $\log\log(16n)$. This is because the signals in $\Theta_2$ are either nondecreasing or nonincreasing.

Second, we emphasize that the minimax rate of $\Theta_k$ is only for the Gaussian observations $X \sim N(\theta^*, \sigma^2 I_n)$. With regard to the upper bound, the assumption of Gaussian errors can be easily relaxed to sub-Gaussian errors. However, the sub-Gaussianity cannot be further relaxed, as illustrated below. Consider the observation $X = \theta^* + Z \in \mathbb{R}^n$. Assume i.i.d. error variables $Z_1, \ldots, Z_n \sim p_\gamma$, where the density function is specified as

$$p_\gamma(x) \propto \exp\bigl(-|x|^\gamma\bigr), \tag{11}$$

for some $\gamma \in (0, 2]$. When $\gamma = 2$, we recover the Gaussian-like (sub-Gaussian) error. For $\gamma \in (0, 2)$, we get a heavier tail than the Gaussian one. The following proposition shows that the sub-Gaussian assumption cannot be relaxed.

PROPOSITION 4.1. *Consider the error distribution* (11) *for some* $\gamma \in (0, 2]$. *For the space* $\Theta_3$, *we have the lower bound*,

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_3} \mathbb{E}\bigl\|\widehat{\theta} - \theta^*\bigr\|^2 \geq c\bigl\{\log(en)\bigr\}^{2/\gamma},$$

*for some universal constant* $c > 0$.

Since the desired minimax rate for $\Theta_3$ is $\sigma^2 \log(en)$, Proposition 4.1 implies that the minimax rate under the Gaussian assumption cannot be achieved unless $\gamma = 2$. In other words, unlike Theorem 2.1, a sub-Gaussian tail is necessary for the result of Theorem 4.1, the second important difference between the two spaces $\Theta_k^{\uparrow}$ and $\Theta_k$.

We end this section with a relatively technical discussion of the difference between models $\Theta_k$ and $\Theta_k^{\uparrow}$. Denoting the estimated change points of $\hat{\theta}(\Theta_k^{\uparrow})$ as $\{\widehat{a}_j\}$. Consider the case where $\theta_i^* = \mu$ for $a \leq i \leq b$ and $a \leq \widehat{a}_{j-1} < \widehat{a}_j \leq b$. The error for $\widehat{a}_{j-1} \leq i \leq \widehat{a}_j$ is $|\overline{X}_{(\widehat{a}_{j-1}:\widehat{a}_j]} - \mu|^2$. For simplicity of discussion, let us suppose $\widehat{a}_j - \widehat{a}_{j-1}$ is of order $n/k$. Without isotonic constraint, few additional structures are exploitable between $\widehat{a}_{j-1}$ and $\widehat{a}_j$, and the optimal fit is shown to suffer an extra logarithmic factor. With isotonic constraint, on the other hand, the two change points $\widehat{a}_{j-1}$ and $\widehat{a}_j$ have an additional constraint:

$$|\overline{X}_{(\widehat{a}_{j-1}:\widehat{a}_j]} - \mu|^2 \leq |\overline{X}_{(a:\widehat{a}_j]} - \mu|^2 \vee |\overline{X}_{(\widehat{a}_{j-1}:b]} - \mu|^2.$$

Now for each term on the right-hand side above, one end point is random and the other is fixed. Therefore, both $|\overline{X}_{(a:\widehat{a}_j]} - \mu|^2$ and $|\overline{X}_{(\widehat{a}_{j-1}:b]} - \mu|^2$ are of order $k \log\log(16n/k)/n$, implied by the asymptotics of partial sum processes (cf. Lemma 5.3).

4.3. *Implications for changepoint detection.* The lower bound result in the paper is strongly related to the problem of determining the "region of detectability" (ROD) in the changepoint detection literature. On one hand, when there are multiple changepoints, the ROD has been established in Arias-Castro, Donoho and Huo (2005), where these authors show that in various settings a signal strength of the order at least $\sqrt{\log(en)/n}$ is necessary for consistent detection. A gap exists when there is only one changepoint.

The result of Theorem 2.2 helps close this gap. As a matter of fact, by a slight modification of the proof of Theorem 2.2 for the case $k = 2$, it is straightforward to prove the following proposition. The result shows that it is impossible to differentiate the one-step function from a two-step function when the signal gap is of order smaller than $\sqrt{\log\log(16n)/n}$. On the other hand, consistent detection of a signal when the gap is of a comparable order has already been established (see, e.g., Chapter 1.5 in Csörgő and Horváth (1997)).

PROPOSITION 4.2. *Let $\mathbb{E}_\theta$ stand for the expectation induced by $N(\theta, \sigma^2 I_n)$. Define the following parameter space*:

$$\Theta_2(c) := \{\theta \in \Theta_2 : (\mu_2 - \mu_1)^2 \cdot (a_1 \wedge (n - a_1)) > c\sigma^2 \log\log(16n)\},$$

*where $\mu_1, \mu_2, a_1$ are defined in* (10). *We then have, for some small enough universal constant $c > 0$,*

$$\inf_{0\le\phi\le 1}\Big\{\sup_{\theta\in\Theta_1}\mathbb{E}_\theta\phi + \sup_{\theta\in\Theta_2(c)}\mathbb{E}_\theta(1-\phi)\Big\} \ge c_1,$$

*where $c_1$ is another universal constant in* $(0, 1)$.

Proposition 4.2 complements Theorem 2.3 in Arias-Castro, Donoho and Huo (2005), and both results together give a clear picture of the ROD when one or multiple changepoints are present.

4.4. *Minimax rates for unimodal piecewise constant functions.* The class of unimodal functions is widely studied in the literature (Bickel and Fan (1996), Birgé (1997), Shoung and Zhang (2001), Köllmann, Bornkamp and Ickstadt (2014)). It is often studied side by side with the isotonic functions (Boyarshinov and Magdon-Ismail (2006), Stout (2008)). In this section, we show that the techniques developed in this paper also lead to the derivation of the minimax rate of the class of unimodal piecewise constant functions. We define the parameter space of interest as follow,

$$\begin{aligned}\Theta_k^{\wedge} = \{\theta \in \mathbb{R}^n : &\text{ there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that}\\ &0 = a_0 \le a_1 \le \cdots \le a_k = n, \mu_1 \le \cdots \le \mu_{\ell-1} \le \mu_\ell \ge \mu_{\ell+1} \ge \cdots \ge \mu_k,\\ &\text{and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j]\}.\end{aligned}$$

This class has been studied by Chatterjee and Lafferty (2019), who provide an upper bound of order $\sigma^2 k \log(en)$. It is interesting to note the relation

$$\Theta_k^{\uparrow} \subset \Theta_k^{\wedge} \subset \Theta_k,$$

which indicates that the minimax rate of $\Theta_k$ is between those of $\Theta_k^{\uparrow}$ and $\Theta_k$. The following theorem gives the exact minimax rate.

THEOREM 4.2. *For any $k \in [n]$, the minimax rate for the space $\Theta_k^{\wedge}$ is given by*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta_k^{\wedge}}\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \asymp \begin{cases}\sigma^2, & k = 1,\\ \sigma^2\log\log(16n), & k = 2,\\ \sigma^2\log(en), & 3 \le k \le \dfrac{\log(en)}{\log\log(16n)},\\ \sigma^2 k\log\log(16n/k), & k > \dfrac{\log(en)}{\log\log(16n)},\end{cases}$$

*where the expectation is taken over the distribution $X \sim N(\theta^*, \sigma^2 I_n)$.*

Interestingly, we observe that the minimax rates have four phases, and can have either a logarithmic behavior or an iterated logarithmic behavior, depending on the regime of $k$. When $k = 2$, the minimax rate is driven by the isotonic structure. When $3 \le k \le \frac{\log(en)}{\log\log(16n)}$, the rate $\sigma^2\log(en)$ results from the uncertainty of the mode of the function. Finally, the $\sigma^2 k\log\log(16n/k)$ rate for a large $k$ is again driven by the isotonic structure of a unimodal function.

4.5. *Minimax rates under $\ell_p$ loss with $1 \le p < 2$.* Section 2 gives the minimax rate of the space $\Theta_k^{\uparrow}$ with respect to the squared $\ell_2$ loss. In particular, Theorems 2.1 and 2.2 show that the minimax rate involves an interesting iterated logarithmic term. This is in contrast with the original isotonic regression estimator $\hat{\theta}(\Theta_n^{\uparrow})$, which is of an additional logarithmic term in view of Proposition 2.1.

In this section we present an interesting phenomenon that a reversed argument applies to an $\ell_p$ loss with $1 \le p < 2$. For this, we first reveal that the difference between the minimax risk of $\Theta_k^{\uparrow}$ and the rate of $\hat{\theta}(\Theta_n^{\uparrow})$ will vanish when we consider an $\ell_p$ loss with $1 \le p < 2$.

PROPOSITION 4.3. *Consider $X = \theta^* + Z$ with $Z_1, \ldots, Z_n$ independent, mean zero, and satisfying $\mathbb{E}|Z_i/\sigma|^2 \le C_1$ for some universal constant $C_1 > 0$. We then have, for any $k \in [n]$ and $1 \le p < 2$,*

$$\sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\hat{\theta}\big(\Theta_n^{\uparrow}\big) - \theta^*\big\|_p^p \le C\sigma^p n(k/n)^{p/2}$$

*for some universal constant $C > 0$. On the other hand, there exists some universal constant $c > 0$ such that, for any $1 \le p < 2$,*

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\hat{\theta} - \theta^*\big\|_p^p \ge c\sigma^p n(k/n)^{p/2},$$

*where the infimum is taken over all measurable functions of $X$ and the expectation is taken under which $X \sim N(\theta^*, \sigma^2 I_n)$.*

Second we show that, quite interestingly, the reduced isotonic regression estimator cannot recover the above minimax risk under an $\ell_p$ loss with $1 \le p < 2$, even if it is the maximum likelihood estimator of the truth.

PROPOSITION 4.4. *Consider $X = \theta^* + Z$ with $Z \sim N(0, \sigma^2 I_n)$. Then, for any $1 \le p \le 2$ and $2 \le k \le n$, we have*

$$\sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\big\|\hat{\theta}\big(\Theta_k^{\uparrow}\big) - \theta^*\big\|_p^p \asymp \sigma^p n\big\{k \log\log(16n/k)/n\big\}^{p/2}.$$

Unlike the estimator $\widehat{\theta}(\Theta_n^{\uparrow})$, for $\widehat{\theta}(\Theta_k^{\uparrow})$, the iterated logarithmic term does not disappear when an $\ell_p$ loss with $1 \le p < 2$ is considered. Since Proposition 4.4 gives both upper and lower bounds for the $\ell_p$ risk, the reduced isotonic regression estimator $\widehat{\theta}(\Theta_k^{\uparrow})$ is not optimal for the class $\Theta_k^{\uparrow}$ when $1 \le p < 2$, compared with the minimax rate given in Proposition 4.3. This indicates that, compared to the classical isotonic regression estimator, the performance of the reduced isotonic regression estimator hinges more on its definition, that is, minimizing the squared $\ell_2$ risk. This interesting phenomenon is summarized in Table 1. The rates displayed are for the normalized $\ell_p$ loss $\|\hat{\theta} - \theta^*\|_p^p/(n\sigma^p)$.

**5. Proofs.** This section contains the proofs of the main results in Sections 2 and 3, with the remaining proofs and auxiliary lemmas relegated to the Supplementary Material (Gao, Han and Zhang (2020)). In the sequel, by convention the summation over an empty set is set to be 0.

TABLE 1
*The minimax rates and the rates of convergence of $\widehat{\theta}(\Theta_k^{\uparrow})$ and $\widehat{\theta}(\Theta_n^{\uparrow})$ for the class $\Theta_k^{\uparrow}$ under the normalized $\ell_p$ loss $\|\hat{\theta}-\theta^*\|_p^p/(n\sigma^p)$*

| | minimax rate | $\widehat{\theta}(\Theta_k^{\uparrow})$ | $\widehat{\theta}(\Theta_n^{\uparrow})$ |
|---|---|---|---|
| $1 \le p < 2$ | $\left(\frac{k}{n}\right)^{p/2}$ | $\left(\frac{k\log\log(16n/k)}{n}\right)^{p/2}$ | $\left(\frac{k}{n}\right)^{p/2}$ |
| $p = 2$ | $\frac{k\log\log(16n/k)}{n}$ | $\frac{k\log\log(16n/k)}{n}$ | $\frac{k\log(en/k)}{n}$ |

5.1. *A critical lemma.* Before stating the proofs of all theorems in the paper, we first present a very important lemma that characterizes the solution $\widehat{\theta}(\Theta_k^{\uparrow})$ of the reduced isotonic regression (1). Below, we use the notation $\widehat{\theta}^{(k)}$ for $\widehat{\theta}(\Theta_k^{\uparrow})$, and recall the set of knots of $\widehat{\theta}^{(k)} = \{\widehat{a}_j\}$ is denoted as $\widehat{A}_k$.

LEMMA 5.1. *The following properties of estimator $\widehat{\theta}^{(k)} = \widehat{\theta}(\Theta_k^{\uparrow})$ hold.*

1. *For each $j$, $\widehat{\theta}_i^{(k)} = \overline{X}_{(\widehat{a}_{j-1}:\widehat{a}_j]}$ for all $i \in (\widehat{a}_{j-1} : \widehat{a}_j]$.*
2. *For each $j$, we have $\overline{X}_{(s:\widehat{a}_j]} < \frac{\widehat{\theta}^{(k)}_{\widehat{a}_j} + \widehat{\theta}^{(k)}_{\widehat{a}_{j+1}}}{2} < \overline{X}_{(\widehat{a}_j:t]}$ for all $0 \le s < \widehat{a}_j < t \le n$. As a consequence, $\widehat{\theta}^{(n)}_{\widehat{a}_j} < \frac{\widehat{\theta}^{(k)}_{\widehat{a}_j} + \widehat{\theta}^{(k)}_{\widehat{a}_{j+1}}}{2} < \widehat{\theta}^{(n)}_{\widehat{a}_j+1}$.*
3. *The set of knots satisfies $\widehat{A}_k \subset \widehat{A}_n$.*

These three results in Lemma 5.1 are all deterministic consequences of the optimization problem (1). The first conclusion asserts that given the set of knots $\widehat{A}_k$, the value of $\widehat{\theta}_i^{(k)}$ is a simple average of $X$ in each block $(\widehat{a}_{j-1} : \widehat{a}_j]$. The second conclusion is due to the isotonic constraint in (1), and is also the reason why we can apply a nonasymptotic law of iterated logarithm bound for the risk (see the Proof of Theorem 2.1). Finally, the last conclusion $\widehat{A}_k \subset \widehat{A}_n$ leads to the efficient computational strategy we discuss in Section 4.1. The proof of the lemma is given below.

PROOF OF LEMMA 5.1. For notational simplicity, we use $\widehat{\theta}$ for $\widehat{\theta}^{(k)} = \widehat{\theta}(\Theta_k^{\uparrow})$ in the proof. We first show that $\widehat{\theta}_i = \overline{X}_{(\widehat{a}_{j-1}:\widehat{a}_j]}$ for all $i \in (\widehat{a}_{j-1} : \widehat{a}_j]$. By the definition of $\Theta_k^{\uparrow}$, the optimization $\min_{\theta\in\Theta_k^{\uparrow}} \|X-\theta\|^2$ can be equivalently written as

$$\min_{a_0\le\cdots\le a_k} \min_{\mu_1\le\cdots\le\mu_k} \sum_{j=1}^{k} \sum_{i=a_{j-1}+1}^{a_j} (X_i-\mu_j)^2$$

$$= \min_{a_0\le\cdots\le a_k} \Bigg\{ \sum_{j=1}^{k} \sum_{i=a_{j-1}+1}^{a_j} (X_i - \overline{X}_{(a_{j-1}:a_j]})^2$$

$$+ \min_{\mu_1\le\cdots\le\mu_k} \sum_{j=1}^{k} (a_j - a_{j-1})(\mu_j - \overline{X}_{(a_{j-1}:a_j]})^2 \Bigg\}.$$

The optimization problem

$$\min_{\mu_1\le\cdots\le\mu_k} \sum_{j=1}^{k} (a_j - a_{j-1})(\mu_j - \overline{X}_{(a_{j-1}:a_j]})^2$$

is in the form of weighted isotonic regression. Therefore, its solution can be represented as

$$\widehat{\mu}_j = \min_{v \geq j} \max_{u \leq j} \overline{X}_{(a_{u-1}:a_v]}. \tag{12}$$

This fact can be derived using the same proof of the minimax formula of isotonic regression (cf. Proposition 2.4.2 in Silvapulle and Sen (2011)). Now suppose $(\widetilde{a}_0, \ldots, \widetilde{a}_k)$ is a minimizer, then the solution has the form $\widehat{\theta}_i = \min_{v \geq j} \max_{u \leq j} \overline{X}_{(\widetilde{a}_{u-1}:\widetilde{a}_v]}$ for all $i \in (\widetilde{a}_{j-1} : \widetilde{a}_j]$. Note that the values in the $k$ intervals satisfy $\widehat{\mu}_1 \leq \cdots \leq \widehat{\mu}_k$. We can combine any two adjacent interval if $\widehat{\mu}_{j-1} = \widehat{\mu}_j$. Then, by the formula (12), there exist $\{\widehat{a}_j\}$ such that $\widehat{\theta}_i = \overline{X}_{(\widehat{a}_{j-1}:\widehat{a}_j]}$ for all $i \in (\widehat{a}_{j-1} : \widehat{a}_j]$.

Now we prove the second point. By symmetry, it is sufficient to prove $(\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2 < \overline{X}_{(\widehat{a}_j:t]}$. Moreover, as $\widehat{\theta}_i$ is nondecreasing in $i$, it suffices to consider $\widehat{a}_j < t < \widehat{a}_{j+1}$. There are three possible cases.

*Case 1.* $\overline{X}_{(t:\widehat{a}_{j+1}]} \neq \widehat{\theta}_{\widehat{a}_{j+1}}$ and $\overline{X}_{(t:\widehat{a}_{j+1}]} \geq \widehat{\theta}_{\widehat{a}_j}$. By the optimality of $\widehat{\theta}(\Theta_k^{\uparrow})$, assigning $\widehat{\theta}_{\widehat{a}_j}$ to $\widehat{\theta}_i$ for all $i \in (\widehat{a}_j : t]$ does not provide a better fit,

$$\sum_{i=\widehat{a}_j+1}^{t} (X_i - \widehat{\theta}_{\widehat{a}_{j+1}})^2 + \sum_{i=t+1}^{\widehat{a}_{j+1}} (X_i - \widehat{\theta}_{\widehat{a}_{j+1}})^2$$

$$\leq \sum_{i=\widehat{a}_j+1}^{t} (X_i - \widehat{\theta}_{\widehat{a}_j})^2 + \sum_{i=t+1}^{\widehat{a}_{j+1}} (X_i - \overline{X}_{(t:\widehat{a}_{j+1}]})^2.$$

It follows that

$$(t - \widehat{a}_j)(\overline{X}_{(\widehat{a}_j:t]} - \widehat{\theta}_{\widehat{a}_{j+1}})^2 + (\widehat{a}_{j+1} - t)(\overline{X}_{(t:\widehat{a}_{j+1}]} - \widehat{\theta}_{\widehat{a}_{j+1}})^2$$

$$\leq (t - \widehat{a}_j)(\overline{X}_{(\widehat{a}_j:t]} - \widehat{\theta}_{\widehat{a}_j})^2.$$

This leads to $|\overline{X}_{(\widehat{a}_j:t]} - \widehat{\theta}_{\widehat{a}_{j+1}}| < |\overline{X}_{(\widehat{a}_j:t]} - \widehat{\theta}_{\widehat{a}_j}|$, which further implies $\overline{X}_{(\widehat{a}_j:t]} > (\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2$.

*Case 2.* $\overline{X}_{(t:\widehat{a}_{j+1}]} = \widehat{\theta}_{\widehat{a}_{j+1}}$. Since $\widehat{\theta}_{\widehat{a}_{j+1}} = \overline{X}_{(\widehat{a}_j:\widehat{a}_{j+1}]}$ is a weighted average of $\overline{X}_{(t:\widehat{a}_{j+1}]}$ and $\overline{X}_{(\widehat{a}_j:t]}$, we have $\overline{X}_{(\widehat{a}_j:t]} = \widehat{\theta}_{\widehat{a}_{j+1}} > \widehat{\theta}_{\widehat{a}_j}$. Thus, we still have $\overline{X}_{(\widehat{a}_j:t]} > (\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2$.

*Case 3.* $\overline{X}_{(t:\widehat{a}_{j+1}]} < \widehat{\theta}_{\widehat{a}_j}$. By the definition of $\{\widehat{a}_j\}$, we have $\widehat{\theta}_{\widehat{a}_{j+1}} > \widehat{\theta}_{\widehat{a}_j}$. Moreover, since $\widehat{\theta}_{\widehat{a}_{j+1}} = \overline{X}_{(\widehat{a}_j:\widehat{a}_{j+1}]}$ is a weighted average of $\overline{X}_{(t:\widehat{a}_{j+1}]}$ and $\overline{X}_{(\widehat{a}_j:t]}$, we must have $\overline{X}_{(\widehat{a}_j:t]} > \widehat{\theta}_{\widehat{a}_{j+1}}$ and $\overline{X}_{(\widehat{a}_j:t]} > \widehat{\theta}_{\widehat{a}_j}$, which also leads to $\overline{X}_{(\widehat{a}_j:t]} > (\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2$.

Finally, we have

$$\widehat{\theta}^{(n)}_{\widehat{a}_j+1} = \min_{b \geq \widehat{a}_j+1} \max_{a \leq \widehat{a}_j+1} \overline{X}_{[a:b]} \geq \min_{b > \widehat{a}_j} \overline{X}_{(\widehat{a}_j:b]} > (\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2.$$

By symmetry, we also have $\widehat{\theta}^{(n)}_{\widehat{a}_j} < (\widehat{\theta}_{\widehat{a}_j} + \widehat{\theta}_{\widehat{a}_{j+1}})/2$, and therefore $\widehat{\theta}^{(n)}_{\widehat{a}_j} < \widehat{\theta}^{(n)}_{\widehat{a}_j+1}$, meaning that $\widehat{a}_j$ is also a change point for $\widehat{\theta}^{(n)}$, which immediately implies the last conclusion $\widehat{A}_k \subset \widehat{A}_n$. □

5.2. *Proofs of upper bounds.* In this section, we state the proofs of Theorems 2.1 and 3.1.

PROOF OF THEOREM 2.1. We first introduce notation that is needed in the proof. We shorthand $\widehat{\theta}(\Theta_k^{\uparrow})$ by $\widehat{\theta}$. The set of knots of $\widehat{\theta}$ is denoted by $\widehat{A}_k = \{\widehat{a}_h\}$. Define the oracle

$$\theta^{(k)} = \operatorname*{argmin}_{\theta \in \Theta_k^{\uparrow}} \|\theta - \theta^*\|^2. \tag{13}$$

The set of knots of $\theta^{(k)}$ is denoted by $A_k = \{a_j\}$ where we allow overlaps within $a_1, \ldots, a_k$. For the error vector $Z = X - \theta^*$ and two integers $1 \le a \le b \le n$, define random variables

$$\xi_+(a,b,\ell) = 2^\ell \max\{|\overline{Z}_{(a:I]}|^2 : a + 2^{\ell-1} \le t \le b \wedge (a + 2^\ell - 1)\}, \tag{14}$$

$$\delta_+(a,b,\ell) = \max_{\{h: a < \widehat{a}_h \le b\}} \mathbb{1}\{a + 2^{\ell-1} \le \widehat{a}_h \le b \wedge (a + 2^\ell - 1)\}, \tag{15}$$

$$\xi_-(a,b,\ell) = 2^\ell \max\{|\overline{Z}_{(t:b]}|^2 : a \vee (b + 2 - 2^\ell) \le t \le b + 1 - 2^{\ell-1}\}, \tag{16}$$

$$\delta_-(a,b,\ell) = \max_{\{h: a < \widehat{a}_h \le b\}} \mathbb{1}\{a \vee (b + 2 - 2^\ell) \le \widehat{a}_h \le b + 1 - 2^{\ell-1}\}. \tag{17}$$

We adopt the convention that maximum over an empty set is zero. The random variables defined above satisfy the following lemma, which will be proved in Section A3 in the Supplementary Material (Gao, Han and Zhang (2020)).

LEMMA 5.2. *There exists a universal constant $C > 0$, such that for any integer $f \ge 0$,*

$$\sum_{j=1}^{k} \sum_{\{\ell \ge 1: a_{j-1} + 2^{\ell-1} \le a_j\}} \mathbb{E}\delta_+(a_{j-1}, a_j, \ell + f)\xi_+(a_{j-1}, a_j, \ell) \le C\sigma^2 k \log\log(16n/k),$$

$$\sum_{j=1}^{k} \sum_{\{\ell \ge 1: a_{j-1} \le a_j - 2^{\ell-1}\}} \mathbb{E}\delta_-(a_{j-1}, a_j, \ell + f)\xi_-(a_{j-1}, a_j, \ell) \le C\sigma^2 k \log\log(16n/k).$$

We also need the following lemma to facilitate the proof. Its proof will also be given in Section A3 in the Supplementary Material (Gao, Han and Zhang (2020)).

LEMMA 5.3. *There exists a universal constant $C > 0$, such that*

$$\sum_{j=1}^{k} \mathbb{E} \max_{a_{j-1} < a \le a_j} (a - a_{j-1})\overline{Z}^2_{(a_{j-1}:a]} \le C\sigma^2 k \log\log(16n/k),$$

$$\sum_{j=1}^{k} \mathbb{E} \max_{a_{j-1} \le a \le a_j} (a_j - a)\overline{Z}^2_{(a:a_j]} \le C\sigma^2 k \log\log(16n/k).$$

The proof of Theorem 2.1 starts with the basic inequality $\|X - \widehat{\theta}\|^2 \le \|X - \theta^{(k)}\|^2$, a direct consequence of the definition of $\widehat{\theta}$. Since

$$\|X - \widehat{\theta}\|^2 = \|X - \theta^*\|^2 + \|\theta^* - \widehat{\theta}\|^2 + 2\langle X - \theta^*, \theta^* - \widehat{\theta}\rangle, \tag{18}$$

$$\|X - \theta^{(k)}\|^2 = \|X - \theta^*\|^2 + \|\theta^* - \theta^{(k)}\|^2 + 2\langle X - \theta^*, \theta^* - \theta^{(k)}\rangle, \tag{19}$$

we have

$$\|\widehat{\theta} - \theta^*\|^2 \le \|\theta^{(k)} - \theta^*\|^2 + 2\langle X - \theta^*, \widehat{\theta} - \theta^{(k)}\rangle. \tag{20}$$

For each $j$, define $h_j = \max\{h : \widehat{a}_h \le a_j\}$. It is easy to see that $\widehat{a}_{h_j} \le a_{j-1}$ if and only if $\widehat{\theta}$ is a constant in the interval $(a_{j-1} : a_j]$. Then, the inner product term above is

$$2\langle X - \theta^*, \widehat{\theta} - \theta^{(k)}\rangle = 2\sum_{j=1}^{k} \mathbb{1}\{\widehat{a}_{h_j} \le a_{j-1}\} \sum_{i \in (a_{j-1}:a_j]} (X_i - \theta_i^*)(\widehat{\theta}_i - \theta_i^{(k)})$$

$$
\begin{aligned}
&+2\sum_{j=1}^{k}\mathbb{1}\{\widehat{a}_{h_j}>a_{j-1}\}\sum_{i\in(a_{j-1}:a_j]}(X_i-\theta_i^*)\bigl(\widehat{\theta}_i-\theta_i^{(k)}\bigr)\\
&=2\sum_{\{j\in[k]:\widehat{a}_{h_j}\le a_{j-1}\}}(a_j-a_{j-1})\overline{Z}_{(a_{j-1}:a_j]}\bigl(\widehat{\theta}_{a_j}-\theta_{a_j}^{(k)}\bigr)\\
&\quad+2\sum_{\{j\in[k]:\widehat{a}_{h_j}>a_{j-1}\}}(a_j-\widehat{a}_{h_j})\overline{Z}_{(\widehat{a}_{h_j}:a_j]}\bigl(\widehat{\theta}_{a_j}-\theta_{a_j}^{(k)}\bigr)\\
&\quad+2\sum_{\{j\in[k]:\widehat{a}_{h_j}>a_{j-1}\}}(\widehat{a}_{h_{j-1}+1}-a_{j-1})\\
&\quad\times\overline{Z}_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]}\bigl(\widehat{\theta}_{\widehat{a}_{h_{j-1}+1}}-\theta_{\widehat{a}_{h_{j-1}+1}}^{(k)}\bigr)\\
&\quad+2\sum_{\{j\in[k]:\widehat{a}_{h_j}>a_{j-1}\}}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\\
&\quad\times(\widehat{a}_h-\widehat{a}_{h-1})\overline{Z}_{(\widehat{a}_{h-1}:\widehat{a}_h]}\bigl(\widehat{\theta}_{\widehat{a}_h}-\theta_{\widehat{a}_h}^{(k)}\bigr).
\end{aligned}
\tag{21}
$$

The summation over an empty set is understood as zero. The inner product $2\langle X-\theta^*,\widehat{\theta}-\theta^{(k)}\rangle$ is bounded by four terms. For the first three terms, we can use Cauchy–Schwarz and, for any $\eta\in(0,1)$, get the bound

$$
\begin{aligned}
&3\eta\bigl\|\widehat{\theta}-\theta^{(k)}\bigr\|^2+\eta^{-1}\sum_{j=1}^{k}(a_j-a_{j-1})\overline{Z}^2_{(a_{j-1}:a_j]}\\
&\quad+\eta^{-1}\sum_{j=1}^{k}(a_j-\widehat{a}_{h_j})\overline{Z}^2_{(\widehat{a}_{h_j}:a_j]}\\
&\quad+\eta^{-1}\sum_{j=1}^{k}(\widehat{a}_{h_{j-1}+1}-a_{j-1})\overline{Z}^2_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]}.
\end{aligned}
\tag{22}
$$

Bounding the fourth term (21) is involved. We need some extra notation. For each $h$ such that $(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]$, define

$$
\begin{aligned}
a'_{h-1}&=\left\lfloor\frac{a_{j-1}+\widehat{a}_h}{2}\right\rfloor,\qquad b'_{h-1}=\widehat{a}_{h-1}\wedge a'_{h-1},\\
a''_h&=\left\lceil\frac{\widehat{a}_{h-1}+a_j}{2}\right\rceil,\qquad b''_h=\widehat{a}_h\vee a''_h.
\end{aligned}
$$

Given any integers $1\le a\le b\le n$, define the random variables

$$
\begin{aligned}
\overline{Z}'_{(a:b]}&=\max_{b'\in(a:b]}\frac{b'-a}{b-a}|\overline{Z}_{(a:b']}|,\\
\overline{Z}''_{(a:b]}&=\max_{a'\in[a:b)}\frac{b-a'}{b-a}|\overline{Z}_{(a':b]}|.
\end{aligned}
$$

By Lemma 5.1, we have $\widehat{\theta}_{\widehat{a}_h}\le\overline{X}_{(\widehat{a}_h:b''_h]}$. Since $\overline{X}_{(\widehat{a}_{h-1}:b''_h]}$ is a weighted average of $\widehat{\theta}_{\widehat{a}_h}=\overline{X}_{(\widehat{a}_{h-1}:\widehat{a}_h]}$ and $\overline{X}_{(\widehat{a}_h:b''_h]}$, we get $\widehat{\theta}_{\widehat{a}_h}\le\overline{X}_{(\widehat{a}_{h-1}:b''_h]}$. With this bound, we have

$$
\widehat{\theta}_{\widehat{a}_h}-\theta_{\widehat{a}_h}^{(k)}\le\overline{X}_{(\widehat{a}_{h-1}:b''_h]}-\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]}+\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]}-\theta_{\widehat{a}_h}^{(k)}
$$

$$
\begin{aligned}
&= \overline{Z}_{(\widehat{a}_{h-1}:b''_h]} + \overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h} \\
&= \frac{a_j - \widehat{a}_{h-1}}{b''_h - \widehat{a}_{h-1}} \overline{Z}_{(\widehat{a}_{h-1}:a_j]} - \frac{a_j - b''_h}{b''_h - \widehat{a}_{h-1}} \overline{Z}_{(b''_h:a_j]} + \overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h} \\
&\le 4\overline{Z}''_{(\widehat{a}_{h-1}:a_j]} + |\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h}|.
\end{aligned}
$$

A symmetric argument also gives

$$
\widehat{\theta}_{\widehat{a}_h} - \theta^{(k)}_{\widehat{a}_h} \ge -4\overline{Z}'_{(a_{j-1}:\widehat{a}_h]} - |\overline{\theta}^*_{(b'_{h-1}:\widehat{a}_h]} - \theta^{(k)}_{\widehat{a}_h}|.
$$

Therefore, we have the inequality

$$
\begin{aligned}
|\widehat{\theta}_{\widehat{a}_h} - \theta^{(k)}_{\widehat{a}_h}| \le{}& 4\big(\overline{Z}'_{(a_{j-1}:\widehat{a}_h]} \vee \overline{Z}''_{(\widehat{a}_{h-1}:a_j]}\big) \\
&+ |\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h}| \vee |\overline{\theta}^*_{(b'_{h-1}:\widehat{a}_h]} - \theta^{(k)}_{\widehat{a}_h}|.
\end{aligned} \tag{23}
$$

Since (21) is a sum of $k$ terms, we can bound each of the term separately.

For each $j \in [k]$, recalling $\overline{Z}_{(\widehat{a}_{h-1}:\widehat{a}_h]} = \widehat{\theta}_{\widehat{a}_h} - \overline{\theta}^*_{(\widehat{a}_{h-1}:\widehat{a}_h]}$,

$$
\begin{aligned}
&\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) \overline{Z}_{(\widehat{a}_{h-1}:\widehat{a}_h]} \big(\widehat{\theta}_{\widehat{a}_h} - \theta^{(k)}_{\widehat{a}_h}\big) \\
&= \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) \big(\widehat{\theta}_{\widehat{a}_h} - \theta^{(k)}_{\widehat{a}_h}\big)^2 \\
&\quad + \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) \big(\theta^{(k)}_{\widehat{a}_h} - \overline{\theta}^*_{(\widehat{a}_{h-1}:\widehat{a}_h]}\big)\big(\widehat{\theta}_{\widehat{a}_h} - \theta^{(k)}_{\widehat{a}_h}\big) \\
&\le 32 \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}|^2 \\
&\quad + 32 \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}|^2 \\
&\quad + 2 \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h}|^2 \qquad (24) \\
&\quad + 2 \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{\theta}^*_{(b'_{h-1}:\widehat{a}_h]} - \theta^{(k)}_{\widehat{a}_h}|^2 \qquad (25) \\
&\quad + \frac{\eta}{2} \|\widehat{\theta} - \theta^{(k)}\|^2_{(a_{j-1}:a_j]} + \frac{1}{2\eta} \|\theta^{(k)} - \theta^*\|^2_{(a_{j-1}:a_j]}.
\end{aligned}
$$

Among the terms in the above bound, we need to further analyze (24) and (25). We have

$$
\begin{aligned}
&\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]} - \theta^{(k)}_{\widehat{a}_h}|^2 \\
&\le \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{\theta}^*_{(\widehat{a}_{h-1}:\widehat{a}_h]} - \theta^{(k)}_{\widehat{a}_h}|^2 \\
&\quad + \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{\theta}^*_{(\widehat{a}_{h-1}:a''_h]} - \theta^{(k)}_{\widehat{a}_h}|^2 \\
&\le \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}} \|\theta^{(k)} - \theta^*\|^2_{(\widehat{a}_{h-1}:\widehat{a}_h]}
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a''_h-\widehat{a}_{h-1}}\sum_{i\in(\widehat{a}_{h-1}:a''_h]}(\theta^*_i-\theta^{(k)}_i)^2\\
&\le\|\theta^{(k)}-\theta^*\|^2_{(a_{j-1}:a_j]}\\
&\quad+\sum_{i\in(a_{j-1}:a_j]}(\theta^*_i-\theta^{(k)}_i)^2\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a''_h-\widehat{a}_{h-1}}\mathbb{1}\{\widehat{a}_{h-1}<i\le a''_h\},
\end{aligned}
$$

where

$$
\begin{aligned}
&\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a''_h-\widehat{a}_{h-1}}\mathbb{1}\{\widehat{a}_{h-1}<i\le a''_h\}\\
&\quad\le 2\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a_j-\widehat{a}_{h-1}}\mathbb{1}\{\widehat{a}_{h-1}<i\le a''_h\}\\
&\quad\le 2\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a_j-i+1}\mathbb{1}\{\widehat{a}_{h-1}<i\le a''_h\}\qquad(26)\\
&\quad\le 2\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\frac{\widehat{a}_h-\widehat{a}_{h-1}}{a_j-i+1}\mathbb{1}\{a_j-\widehat{a}_{h-1}\le 2(a_j-i+1)\}\\
&\quad\le 2\max\left\{\frac{a_j-\widehat{a}_t}{a_j-i+1}:a_j-\widehat{a}_t\le 2(a_j-i+1)\right\}\le 4.
\end{aligned}
$$

The inequality (26) above is due to the fact that $i\le a''_h\le\frac{\widehat{a}_{h-1}+a_j+1}{2}$ implies

$$
2(a_j-i+1)\ge 2a_j+2-(\widehat{a}_{h-1}+a_j+1)=a_j-\widehat{a}_{h-1}+1.
$$

Therefore, we obtain

$$
\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})|\overline{\theta}^*_{(\widehat{a}_{h-1}:b''_h]}-\theta^{(k)}_{\widehat{a}_h}|^2\le 5\|\theta^{(k)}-\theta^*\|^2_{(a_{j-1}:a_j]},
$$

which leads to a bound for (24). A symmetric argument gives

$$
\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})|\overline{\theta}^*_{(b'_{h-1}:\widehat{a}_h]}-\theta^{(k)}_{\widehat{a}_h}|^2\le 5\|\theta^{(k)}-\theta^*\|^2_{(a_{j-1}:a_j]},
$$

which leads to a bound for (25). Summing over $j\in[k]$, a bound for (21) is given by

$$
\begin{aligned}
&64\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})|\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}|^2\\
&\quad+64\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})|\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}|^2\\
&\quad+(40+\eta^{-1})\|\theta^{(k)}-\theta^*\|^2+\eta\|\widehat{\theta}-\theta^{(k)}\|^2.
\end{aligned}
$$

We can plug the above bound and (22) into (20), and we get

$$
\begin{aligned}
\|\widehat{\theta}-\theta^*\|^2&\le(41+\eta^{-1})\|\theta^{(k)}-\theta^*\|^2+4\eta\|\widehat{\theta}-\theta^{(k)}\|^2\\
&\quad+\eta^{-1}\sum_{j=1}^{k}(a_j-a_{j-1})\overline{Z}^2_{(a_{j-1}:a_j]}+\eta^{-1}\sum_{j=1}^{k}(a_j-\widehat{a}_{h_j})\overline{Z}^2_{(\widehat{a}_{h_j}:a_j]}
\end{aligned}
$$

$$+\eta^{-1}\sum_{j=1}^{k}(\widehat{a}_{h_{j-1}+1}-a_{j-1})\overline{Z}^2_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]}$$

$$+64\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})\bigl|\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}\bigr|^2$$

$$+64\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})\bigl|\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}\bigr|^2.$$

Use the inequality $\|\widehat{\theta}-\theta^{(k)}\|^2\le 2\|\widehat{\theta}-\theta^*\|^2+2\|\theta^{(k)}-\theta^*\|^2$, set $\eta=1/16$, and some rearrangement of the above bound gives

$$\bigl\|\widehat{\theta}-\theta^*\bigr\|^2\le C\bigl\|\theta^{(k)}-\theta^*\bigr\|^2+C\sum_{j=1}^{k}(a_j-a_{j-1})\overline{Z}^2_{(a_{j-1}:a_j]}$$

$$+C\sum_{j=1}^{k}(\widehat{a}_{h_{j-1}+1}-a_{j-1})\overline{Z}^2_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]}+C\sum_{j=1}^{k}(a_j-\widehat{a}_{h_j})\overline{Z}^2_{(\widehat{a}_{h_j}:a_j]}$$

$$+C\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})\bigl|\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}\bigr|^2$$

$$+C\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})\bigl|\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}\bigr|^2,$$

where $C>0$ is some universal constant. Note that

$$\mathbb{E}\sum_{j=1}^{k}(a_j-a_{j-1})\overline{Z}^2_{(a_{j-1}:a_j]}=k\sigma^2\lesssim\sigma^2k\log\log(16n/k).$$

By Lemma 5.2, we have

$$\mathbb{E}\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})\bigl|\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}\bigr|^2$$

$$=\mathbb{E}\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h-\widehat{a}_{h-1})$$

$$\times\max_{b\in(a_{j-1}:\widehat{a}_h]}\frac{(b-a_{j-1})^2}{(\widehat{a}_h-a_{j-1})^2}|\overline{Z}_{(a_{j-1}:b]}|^2$$

$$\le\mathbb{E}\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\sum_{\{\ell\ge1:a_{j-1}+2^{\ell-1}\le a_j\}}\mathbb{1}\{a_{j-1}+2^{\ell-1}$$

$$\le\widehat{a}_h<a_{j-1}+2^{\ell}\}$$

$$\times(\widehat{a}_h-\widehat{a}_{h-1})\max_{b\in(a_{j-1}:\widehat{a}_h]}\frac{(b-a_{j-1})^2}{(\widehat{a}_h-a_{j-1})^2}|\overline{Z}_{(a_{j-1}:b]}|^2$$

$$\le\mathbb{E}\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}\sum_{\{\ell\ge1:a_{j-1}+2^{\ell-1}\le a_j\}}\mathbb{1}\{a_{j-1}+2^{\ell-1}$$

$$
\begin{aligned}
&\le \widehat{a}_h < a_{j-1} + 2^\ell\big\} \\
&\quad \times (\widehat{a}_h - \widehat{a}_{h-1}) \max_{b \in (a_{j-1}:a_j \wedge (a_{j-1}+2^\ell-1)]} \frac{(b-a_{j-1})^2}{(2^{\ell-1})^2} |\overline{Z}_{(a_{j-1}:b]}|^2 \\
&\le \mathbb{E} \sum_{j=1}^k \sum_{\{\ell \ge 1: a_{j-1}+2^{\ell-1} \le a_j\}} \delta_+(a_{j-1}, a_j, \ell) 2^\ell \\
&\quad \times \max_{b \in (a_{j-1}:a_j \wedge (a_{j-1}+2^\ell-1)]} \frac{(b-a_{j-1})^2}{(2^{\ell-1})^2} |\overline{Z}_{(a_{j-1}:b]}|^2 \\
&\le 4\mathbb{E} \sum_{j=1}^k \sum_{\{\ell \ge 1: a_{j-1}+2^{\ell-1} \le a_j\}} \delta_+(a_{j-1}, a_j, \ell) 2^{-\ell} \\
&\quad \times \max_{b \in (a_{j-1}:a_j \wedge (a_{j-1}+2^\ell-1)]} (b-a_{j-1})^2 |\overline{Z}_{(a_{j-1}:b]}|^2 \\
&\le 4\mathbb{E} \sum_{j=1}^k \sum_{\{\ell \ge 1: a_{j-1}+2^{\ell-1} \le a_j\}} \delta_+(a_{j-1}, a_j, \ell) 2^{-\ell} \sum_{f=1}^\ell 2^f \xi_+(a_{j-1}, a_j, f) \\
&\le 4\mathbb{E} \sum_{j=1}^k \sum_{f \ge 0} 2^{-f} \sum_{\{\ell \ge 1: a_{j-1}+2^{\ell-1} \le a_j\}} \delta_+(a_{j-1}, a_j, \ell+f) \xi_+(a_{j-1}, a_j, \ell),
\end{aligned}
$$

which leads to the conclusion

$$
\begin{aligned}
&\mathbb{E} \sum_{j=1}^k \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h] \subset (a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}|^2 \\
&= 4 \sum_{f \ge 0} 2^{-f} \sum_{j=1}^k \sum_{\{\ell \ge 1: a_{j-1}+2^{\ell-1} \le a_j\}} \mathbb{E} \delta_+(a_{j-1}, a_j, \ell+f) \xi_+(a_{j-1}, a_j, \ell) \qquad (27) \\
&\lesssim \sigma^2 k \log\log(16n/k).
\end{aligned}
$$

Similarly, we also have

$$
\mathbb{E} \sum_{j=1}^k \sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h] \subset (a_{j-1}:a_j]\}} (\widehat{a}_h - \widehat{a}_{h-1}) |\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}|^2 \lesssim \sigma^2 k \log\log(16n/k).
$$

Finally, by Lemma 5.3, we have

$$
\begin{aligned}
&\mathbb{E} \sum_{j=1}^k (\widehat{a}_{h_{j-1}+1} - a_{j-1}) \overline{Z}^2_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]} + \mathbb{E} \sum_{j=1}^k (a_j - \widehat{a}_{h_j}) \overline{Z}^2_{(\widehat{a}_{h_j}:a_j]} \\
&\lesssim \sigma^2 k \log\log(16n/k).
\end{aligned}
$$

Combining the above bounds, we obtain the desired oracle inequality as long as $k \ge 2$.

To complete the proof, we also give the argument for $k = 1$. In this case $\widehat{\theta}_i = \overline{X}$ and $\theta_i^{(1)} = \overline{\theta}^*$ for all $i \in [n]$. Therefore, $\mathbb{E}\|\widehat{\theta} - \theta^{(1)}\|^2 = \sigma^2$, which leads to $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \le 2\|\theta^{(1)} - \theta^*\|^2 + 2\sigma^2$. $\square$

PROOF OF THEOREM 3.1. We use the same notation in the proof of Theorem 2.1, except that $\widehat{\theta}$ is now for $\widehat{\theta}(\Theta^{\uparrow}_{\widehat{k}})$ and $\widehat{A}_{\widehat{k}} = \{\widehat{a}_h\}$. By the definition of $\widehat{\theta}$, we have

$$\|X - \widehat{\theta}\|^2 + \operatorname{pen}_\tau(\widehat{k}) \le \|X - \widehat{\theta}(\Theta^{\uparrow}_k)\|^2 + \operatorname{pen}_\tau(k) \le \|X - \theta^{(k)}\|^2 + \operatorname{pen}_\tau(k).$$

By (18) and (19), we obtain the following inequality:

$$\|\widehat{\theta} - \theta^*\|^2 + \operatorname{pen}_\tau(\widehat{k}) \le \|\theta^{(k)} - \theta^*\|^2 + 2\langle X - \theta^*, \widehat{\theta} - \theta^{(k)}\rangle + \operatorname{pen}_\tau(k). \tag{28}$$

After bounding $2\langle X - \theta^*, \widehat{\theta} - \theta^{(k)}\rangle$ by the same argument in the proof of Theorem 2.1, we obtain

$$\begin{aligned}
&\|\widehat{\theta} - \theta^*\|^2 + 2\operatorname{pen}_\tau(\widehat{k}) - 2\operatorname{pen}_\tau(k) \\
&\le C\|\theta^{(k)} - \theta^*\|^2 + C\sum_{j=1}^{k}(a_j - a_{j-1})\overline{Z}^2_{(a_{j-1}:a_j]} && (29)\\
&\quad + C\sum_{j=1}^{k}(\widehat{a}_{h_{j-1}+1} - a_{j-1})\overline{Z}^2_{(a_{j-1}:\widehat{a}_{h_{j-1}+1}]} + C\sum_{j=1}^{k}(a_j - \widehat{a}_{h_j})\overline{Z}^2_{(\widehat{a}_{h_j}:a_j]} && (30)\\
&\quad + C\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h - \widehat{a}_{h-1})\big|\overline{Z}'_{(a_{j-1}:\widehat{a}_h]}\big|^2 && (31)\\
&\quad + C\sum_{j=1}^{k}\sum_{\{h:(\widehat{a}_{h-1}:\widehat{a}_h]\subset(a_{j-1}:a_j]\}}(\widehat{a}_h - \widehat{a}_{h-1})\big|\overline{Z}''_{(\widehat{a}_{h-1}:a_j]}\big|^2, && (32)
\end{aligned}$$

where $C > 0$ is some universal constant. Take expectation on both sides of the inequality, and then the right-hand side can all be bounded similarly as in the proof of Theorem 2.1 except for (31) and (32). In fact, (31) and (32) can be bounded by the same argument that leads to (27). The only difference is that now the $\{\widehat{a}_h\}$ in the definitions of $\delta_+(a_{j-1}, a_j, \ell)$ and $\delta_-(a_{j-1}, a_j, \ell)$ are from $\widehat{A}_{\widehat{k}}$ instead of $\widehat{A}_k$. Therefore, we need the following lemma, whose proof will be given in Section A3 in the Supplementary Material (Gao, Han and Zhang (2020)).

LEMMA 5.4. *There exists a universal constant $C > 0$, such that*

$$\begin{aligned}
&\max\Bigg\{\sum_{f\ge0}2^{-f}\sum_{j=1}^{k}\sum_{\{\ell\ge1:a_{j-1}+2^{\ell-1}\le a_j\}}\mathbb{E}\delta_+(a_{j-1}, a_j, \ell+f)\xi_+(a_{j-1}, a_j, \ell),\\
&\qquad\sum_{f\ge0}2^{-f}\sum_{j=1}^{k}\sum_{\{\ell\ge1:a_{j-1}\le a_j-2^{\ell-1}\}}\mathbb{E}\delta_-(a_{j-1}, a_j, \ell+f)\xi_-(a_{j-1}, a_j, \ell)\Bigg\}\\
&\le C\left\{\sigma^2 k\log\log\left(\frac{16n}{k}\right) + \sigma^2\mathbb{E}\widehat{k}\log\log\left(\frac{16n}{\widehat{k}}\right)\right\},
\end{aligned}$$

*where the $\{\widehat{a}_h\}$ in the definitions of $\delta_+(a_{j-1}, a_j, \ell)$ and $\delta_-(a_{j-1}, a_j, \ell)$ are from $\widehat{A}_{\widehat{k}}$ instead of $\widehat{A}_k$.*

Then, for some (possibly different) universal constant $C > 0$, we have

$$\begin{aligned}
&\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 + 2\mathbb{E}\operatorname{pen}_\tau(\widehat{k})\\
&\le C\|\theta^{(k)} - \theta^*\|^2 + 2\operatorname{pen}_\tau(k)\\
&\quad + C\left\{\sigma^2 k\log\log\left(\frac{16n}{k}\right) + \sigma^2\mathbb{E}\widehat{k}\log\log\left(\frac{16n}{\widehat{k}}\right)\right\}.
\end{aligned}$$

Choosing $\tau = C_1\sigma^2$ with a sufficiently large constant $C_1 > 0$, we get

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \lesssim \|\theta^{(k)} - \theta^*\|^2 + \sigma^2 k \log\log\left(\frac{16n}{k}\right),$$

which is the desired results for $k \geq 2$.

To complete the proof, we also need to give the analysis for $k = 1$. It is easy to see that in this case the bounds in Lemma 5.3 and Lemma 5.4 can be improved to $C\sigma^2$ and $C\sigma^2 + C\sigma^2\mathbb{E}\widehat{k}\log\log(\frac{16n}{k})\mathbb{1}\{\widehat{k} \geq 2\}$, respectively. Therefore, the choice $\tau = C_1\sigma^2$ with a large constant $C_1 > 0$ leads to

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \lesssim \|\theta^{(1)} - \theta^*\|^2 + \sigma^2. \tag{33}$$

The proof is thus complete. □

5.3. *Proofs of lower bounds.* This section is devoted to proving the lower bounds in Section 2.

PROOF OF PROPOSITION 2.1. Without loss of generality, consider the case when $n/k$ is an integer. Then, $[n] = \bigcup_{j=1}^k \mathcal{C}_j$, where $\mathcal{C}_j$ is the $j$th consecutive interval with cardinality $n/k$. Then, we take $\theta^* \in \Theta_k^{\uparrow}$ with $\theta_i^* = \mu_j$ if $i \in \mathcal{C}_j$. Use the notation $\mathcal{H}_n = \{\theta \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n\}$. Then, as long as $\mu_1, \ldots, \mu_k$ are sufficiently separated,

$$\min_{\theta\in\mathcal{H}_n} \sum_{i=1}^n (X_i - \theta_i)^2 = \sum_{j=1}^k \min_{\theta\in\mathcal{H}_{n/k}} \sum_{i\in\mathcal{C}_j} (X_i - \theta_i)^2,$$

with high probability. This high-probability event is denoted as $E$. We take $\mu_j = \kappa j$ for some $\kappa > 0$. Then, as $\kappa \to \infty$, $\mathbb{P}(E^c)$ converges to 0. In other words, $\mathbb{P}(E^c)$ is arbitrarily small for sufficiently large $\kappa$. We have

$$\mathbb{E}\|\hat{\theta} - \theta^*\|^2 \geq \sum_{j=1}^k \mathbb{E}\|\hat{\theta}_{\mathcal{C}_j} - \theta^*_{\mathcal{C}_j}\|^2 - \mathbb{E}\|\hat{\theta} - \theta^*\|^2 \mathbb{1}_{E^c}.$$

Since $\mathbb{E}\|\hat{\theta} - \theta^*\|^2\mathbb{1}_{E^c} \leq \sqrt{\mathbb{E}\|\hat{\theta} - \theta^*\|^4}\sqrt{\mathbb{P}(E^c)}$ is arbitrarily small for sufficiently large $\kappa$, the term $\mathbb{E}\|\hat{\theta} - \theta^*\|^2\mathbb{1}_{E^c}$ can be neglected. It is sufficient to give a lower bound for $\sum_{j=1}^k \mathbb{E}\|\hat{\theta}_{\mathcal{C}_j} - \theta^*_{\mathcal{C}_j}\|^2$. Note that

$$\sum_{j=1}^k \mathbb{E}\|\hat{\theta}_{\mathcal{C}_j} - \theta^*_{\mathcal{C}_j}\|^2 = \sum_{j=1}^k \mathbb{E}\|\Pi_{\mathcal{H}_{n/k}} Z_{\mathcal{C}_j}\|^2,$$

where $\Pi_{\mathcal{H}_{n/k}}$ is the projection operator onto the space $\mathcal{H}_{n/k}$. By Amelunxen et al. (2014), $\|\Pi_{\mathcal{H}_{n/k}} Z_{\mathcal{C}_j}\|^2 \geq C\log(en/k)$, leading to the desired result. □

We continue to state the proofs of other results. The main tool we will use is Fano's lemma. For any probability measures $\mathbb{P}$, $\mathbb{Q}$, define the Kullback–Leibler divergence to be

$$D(\mathbb{P}\|\mathbb{Q}) = \int \left(\log\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}.$$

The Fano's lemma is stated as follows. See Ibragimov and Has' Minskii (2013) and Tsybakov (2009) for references.

PROPOSITION 5.1. *Let $(\Theta, \rho)$ be a metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability measures. For any totally bounded $T \subset \Theta$, define the Kullback–Leibler diameter by*

$$d_{\mathrm{KL}}(T) = \sup_{\theta,\theta' \in T} D(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}).$$

*Then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left[ \rho^2\{\widehat{\theta}(X), \theta\} \geq \frac{\epsilon^2}{4} \right] \geq 1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(\epsilon, T, \rho)}, \tag{34}$$

*for any $\epsilon > 0$, where $\mathcal{M}(\epsilon, T, \rho)$ stands for the packing number of $T$ with radius $\epsilon$ with respect to the metric $\rho$.*

PROOF OF THEOREM 2.2. We only need to deal with the case when $n > C$ for a sufficiently large constant, since when $n \leq C$, the rate is a constant and the conclusion automatically holds.

When $k = 1$, the standard lower bound argument for the one-dimensional normal mean problem (Lehmann and Casella (1998)) applies here, and we get the desired rate.

The case $k = 2$ is studied in Section 2. Combining (34), (4), and (5) gives

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_2} \mathbb{P}\left( \|\widehat{\theta} - \theta\|^2 \geq \frac{\alpha\sigma^2}{80} \log\log_2 n \right) \geq 1 - \frac{6\alpha \log\log_2 n + \log 2}{\log\log_2 n} \geq c$$

with $\alpha = 1/60$ and a sufficiently small value $c > 0$. Thus, with an application of Markov's inequality, we obtain the desired minimax lower bound in expectation.

Now we derive the lower bound for $k \geq 3$.

We first consider the case $n > C$, $k > C$ and $n/k > C$ for some sufficiently large constant $C > 0$. Define the space $\Theta_2^{\uparrow}(\widetilde{n}, a, b) \subset \mathbb{R}^{\widetilde{n}}$ to be the class of vectors of length $\widetilde{n}$ that have two nondecreasing pieces taking values between $a$ and $b$ respectively. Then, construct the following space:

$$\widetilde{T} = \bigtimes_{\ell=1}^{\lceil \frac{k}{2} \rceil} \widetilde{T}_\ell,$$

where for $1 \leq \ell \leq \lceil \frac{k}{2} \rceil - 1$, we define

$$\widetilde{T}_\ell = \Theta_2^{\uparrow}\left\{ \left\lfloor \frac{2n}{k} \right\rfloor, (2\ell - 2)\sqrt{2\alpha\sigma^2 \log\log_2 n}, (2\ell - 1)\sqrt{2\alpha\sigma^2 \log\log_2 n} \right\}$$

and

$$\widetilde{T}_{\lceil \frac{k}{2} \rceil} = \{k\sqrt{2\alpha\sigma^2 \log\log_2 n}\}^{n - \lfloor \frac{2n}{k} \rfloor (\lceil \frac{k}{2} \rceil - 1)}.$$

Observe that $\widetilde{T} \subset \Theta_k^{\uparrow}$. Thus,

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta\|^2 &\geq \inf_{\widehat{\theta}} \sup_{\theta \in \widetilde{T}} \mathbb{E}\|\widehat{\theta} - \theta\|^2 \\ &= \inf_{\widehat{\theta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_{\lceil k/2 \rceil})} \sum_{\ell=1}^{\lceil \frac{k}{2} \rceil} \sup_{\eta_\ell \in \widetilde{T}_\ell} \mathbb{E}\|\widehat{\eta}_\ell - \eta_\ell\|^2 \qquad (35) \\ &\geq \sum_{\ell=1}^{\lceil \frac{k}{2} \rceil - 1} \inf_{\widehat{\eta}_\ell} \sup_{\eta_\ell \in \widetilde{T}_\ell} \mathbb{E}\|\widehat{\eta}_\ell - \eta_\ell\|^2 \end{aligned}$$

$$\geq c_1\left(\left\lceil \frac{k}{2} \right\rceil - 1\right) \log\log \left\lfloor \frac{2n}{k} \right\rfloor \tag{36}$$

$$\geq c_2 k \log\log\left(\frac{16n}{k}\right),$$

where the equality (35) is by taking advantage of the separable structure and a sufficiency argument, and the inequality (36) is by the same argument that we use to derive the lower bound for the case $k = 2$.

Second, we consider the rest of settings. When $n \leq C$, the rate is a constant and the result automatically holds. When $3 \leq k \leq C$, the rate $\log\log 16n$ is immediately a lower bound by the fact that $\Theta_2^{\uparrow} \subset \Theta_k^{\uparrow}$. When $n/k \leq C$, we have $\Theta_{n/C}^{\uparrow} \subset \Theta_k^{\uparrow}$. Therefore,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_k^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta\|^2 \geq \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{n/C}^{\uparrow}} \mathbb{E}\|\widehat{\theta} - \theta\|^2 \geq c_3 n.$$

Hence, the proof is complete. □

**Acknowledgements.** The authors thank Qiyang Han for carefully reading the manuscript and many insightful suggestions and Antoine Picard for pointing out an error in the proof. The authors also thank two referees and an Associate Editor for their helpful feedbacks that greatly improve the paper.

The first author was supported in part by NSF Grant DMS-1712957.

The second author was supported in part by NSF Grant DMS-1712536.

The third author was supported in part by NSF Grants DMS-1513378, IIS-1407939, DMS-1721495 and IIS-1741390.

## SUPPLEMENTARY MATERIAL

**Supplement to "On estimation of isotonic piecewise constant signals"** (DOI: 10.1214/18-AOS1792SUPP; .pdf). This supplement contains proofs of remaining results in Section 4, as well as some auxiliary lemmas.

## REFERENCES

AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369 https://doi.org/10.1109/TIT.2005.850056

BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

BELLEC, P. C. and TSYBAKOV, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* **16** 1879–1892. MR3417801

BICKEL, P. J. and FAN, J. (1996). Some problems on the estimation of unimodal densities. *Statist. Sinica* **6** 23–45. MR1379047

BIRGÉ, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25** 970–981. MR1447736 https://doi.org/10.1214/aos/1069362733

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719 https://doi.org/10.1007/BF01199316

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

BOYARSHINOV, V. and MAGDON-ISMAIL, M. (2006). Linear time isotonic and unimodal regression in the $L_1$ and $L_\infty$ norms. *J. Discrete Algorithms* **4** 676–691. MR2577688 https://doi.org/10.1016/j.jda.2005.07.001

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348 https://doi.org/10.1214/07-AOS558

CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

CHATTERJEE, S. and LAFFERTY, J. (2019). Adaptive risk bounds in unimodal regression. *Bernoulli* **25** 1–25. MR3892309 https://doi.org/10.3150/16-bej922

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2743035

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over $l_p$-balls for $l_q$-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886 https://doi.org/10.1007/BF01199026

FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized $M$-estimation: Fast computation. *J. Comput. Graph. Statist.* **17** 201–224. MR2424802 https://doi.org/10.1198/106186008X285591

GAO, C., HAN, F. and ZHANG, C.-H (2020). Supplement to "On estimation of isotonic piecewise constant signals." https://doi.org/10.1214/18-AOS1792SUPP.

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. MR1180321 https://doi.org/10.1007/978-3-0348-8621-5

HAIMINEN, N., GIONIS, A. and LAASONEN, K. (2008). Algorithms for unimodal segmentation with applications to unimodality detection. *Knowl. Inf. Syst.* **14** 39–57.

HAN, Q. and WELLNER, J. A. (2016). Multivariate convex regression: Global risk bounds and adaptation. Available at arXiv:1601.06844.

IBRAGIMOV, I. A. and HAS' MINSKII, R. Z. (2013). *Statistical Estimation*: *Asymptotic Theory*. Springer.

JEWELL, S. and WITTEN, D. (2018). Exact spike train inference via $\ell_0$ optimization. *Ann. Appl. Stat.* **12** 2457–2482. MR3875708 https://doi.org/10.1214/18-AOAS1162

KIM, A. K. H., GUNTUBOYINA, A. and SAMWORTH, R. J. (2018). Adaptation in log-concave density estimation. *Ann. Statist.* **46** 2279–2306. MR3845018 https://doi.org/10.1214/17-AOS1619

KÖLLMANN, C., BORNKAMP, B. and ICKSTADT, K. (2014). Unimodal regression using Bernstein–Schoenberg splines and penalties. *Biometrics* **70** 783–793. MR3295739 https://doi.org/10.1111/biom.12193

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

LI, H., MUNK, A. and SIELING, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.* **10** 918–959. MR3486421 https://doi.org/10.1214/16-EJS1131

MAIR, P., HORNIK, K. and DE LEEUW, J. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *J. Stat. Softw.* **32** 1–24.

MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920 https://doi.org/10.1214/aos/1015956708

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134 https://doi.org/10.1214/12-STS393

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

SALANTI, G. and ULM, K. (2003). A nonparametric changepoint model for stratifying continuous variables under order restrictions and binary outcome. *Stat. Methods Med. Res.* **12** 351–367. MR2011212 https://doi.org/10.1191/0962280203sm338ra

SCHELL, M. J. and SINGH, B. (1997). The reduced monotonic regression method. *J. Amer. Statist. Assoc.* **92** 128–135.

SHOUNG, J.-M. and ZHANG, C.-H. (2001). Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.* **29** 648–665. MR1865335 https://doi.org/10.1214/aos/1009210684

SILVAPULLE, M. J. and SEN, P. K. (2011). *Constrained Statistical Inference*: *Order, Inequality, and Shape Constraints*. Wiley.

STOUT, Q. F. (2008). Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.* **53** 289–297. MR2649085 https://doi.org/10.1016/j.csda.2008.08.005

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898 https://doi.org/10.1214/aos/1021379864